Algorithmique des réseaux sociaux **2013/2014**

Cours 6 — 25 Novembre

Enseignant: Marc Lelarge *Scribe: Vincent Vidal*

Pour information

– Page web du cours `http://www.di.ens.fr/~lelarge/soc.html`

6.1 Agrégation de classement (2)

Introduction

On a vu que si l'on essaye de définir les axiomes que doivent vérifier les fonctions d'agrégation de permutations, on trouve des contraintes trop fortes pour qu'il existe une fonction d'agrégation les vérifiant toutes. On peut alors tenter d'agréger des scores qui sont plus souples que les permutations.

Les scores correpondent à des permutations incomplètes. En se donnant une règle pour régler les cas où plusieurs scores sont identiques, on peut déduire une permutation de la fonction de score et donc le classement voulu.

Il s'agit donc ici de classement avec un nombre différent de votes par candidat.
exemple :

– *IMDb avec le « Top 250 » des films*
– *Beeradvocate avec le « Top 250 » des bières*

Prenons un exemple. On souhaite acheter un téléphone.

– Le premier téléphone est noté 5/5 avec 2 personnes votantes.
– Le second est noté 4/5 avec 200 personnes votantes.

Il n'est pas évident que le meilleur candidat correponde à celui avec la meilleure moyenne.

Dans le cas de ImdB et de Beeradvocate, la formule utilisée pour le score final est la suivante :

$$\text{weighted ranking}\,(i) = \frac{R_i v_i + Cm}{v_i + m}$$

Où :

– R_i est la note moyenne du candidat i
– v_i est le nombre de votes pour le candidat i
– C est la note moyenne globale de tous les candidats
– m est le nombre minimum de votes que doit avoir un candidat.

6.1.1 Estimation Bayésienne

Formalisation

On considère ici le problème suivant :
On dispose d'une pièce dont on ne connait pas la probabilité de tomber sur Pile.
On lance n fois la pièce et elle tombe s fois sur Pile.
Question : Quelle est la probabilité qu'elle tombe sur Pile au $n+1$-ième lancer ?
Une estimation maximisant la vraisemblance donne : s/n.

On note $\mathbf{\Theta}$ la variable aléatoire représentant la probabilité que la pièce tombe sur Pile et π la densité de probabilité qui lui est associée.

On notera de plus X_i la variable aléatoire correspondant au i-ième lancer ($X_i = 1$ si la pièce est tombée sur Pile). Et $X = \sum_i X_i$ le nombre de Pile.

On a ainsi :

$$\mathbb{P}\left(X = s \,|\, \mathbf{\Theta} = \theta\right) = \binom{n}{s} \theta^s \left(1-\theta\right)^{n-s}$$

La formule de Bayes donne :

$$\mathbb{P}\left(X = s\right) = \int_0^1 \mathbb{P}\left(X = s \,|\, \mathbf{\Theta} = \theta\right) \pi\left(\theta\right) \,\mathrm{d}\theta$$

$$\mathbb{P}\left(a \leq \mathbf{\Theta} \leq b,\, X = s\right) = \int_a^b \mathbb{P}\left(X = s \,|\, \mathbf{\Theta} = \theta\right) \pi\left(\theta\right) \,\mathrm{d}\theta$$

On obtient alors :

$$\mathbb{P}\left(a \leq \mathbf{\Theta} \leq b \,|\, X = s\right) = \frac{\int_a^b \binom{n}{s} \theta^s (1-\theta)^{n-s} \pi\left(\theta\right) \,\mathrm{d}\theta}{\int_0^1 \binom{n}{s} \theta^s (1-\theta)^{n-s} \pi\left(\theta\right) \,\mathrm{d}\theta}$$

Densité homogène

On prend ici une distribution π uniforme sur $[0,\, 1]$, à défaut d'avoir plus d'information.
En considérant la fonction *bêta* définie par :

$$B\left(i,\, j\right) = \frac{(i-1)!\,(j-1)!}{(i+j-1)!} = \int_0^1 \theta^{i-1} \left(1-\theta\right)^{j-1} \,\mathrm{d}\theta$$

On peut écrire :

$$\mathbb{P}\left(a \leq \mathbf{\Theta} \leq b \,|\, X = s\right) = \int_a^b \frac{\theta^s \left(1-\theta\right)^{n-s}}{B\left(s+1,\, n-s+1\right)} \,\mathrm{d}\theta$$

On trouve ainsi la densité de probabilité associée à la variable $\mathbf{\Theta}|X = s$. Sa moyenne s'écrit alors :

$$\mathbb{E}\left[\mathbf{\Theta} \mid X = s\right] = \int_a^b \theta . \frac{\theta^s (1-\theta)^{n-s}}{B(s+1,\, n-s+1)} \,\mathrm{d}\theta = \frac{B(s+2,\, n-s+1)}{B(s+1,\, n-s+1)} = \frac{s+1}{n+2}$$

Densité gaussienne

Nous donnons ici une interprétation à la formule IMDb.

On prend ici une distribution π gaussienne : $\pi \sim \mathcal{N}\left(\mu_0,\, \sigma_0^2\right)$ (on notera $\tau_0 = 1/\sigma_0^2$ la précision, i.e. l'inverse de la variance), et des X_i suivant une loi normale $\mathcal{N}(\theta,\, \sigma^2)$ (de même $\tau = 1/\sigma^2$)

La densité du vecteur $(\mathbf{\Theta},\, X_1,\, \ldots,\, X_n)$ est donnée par :

$$\sqrt{\frac{\tau_0}{2\pi}} \left(\sqrt{\frac{\tau}{2\pi}}\right)^n \exp\left(-\tau_0 \frac{(\theta-\mu_0)^2}{2} - \tau \sum_i \frac{(X_i-\theta)^2}{2}\right)$$

On souhaite trouver la densité de probabilité $\pi(\theta\,|X_1,\, \ldots,\, X_n)$. On a alors :

$$\begin{aligned} \pi(\theta\,|X_1,\, \ldots,\, X_n) &\propto \exp\left(-\frac{\tau_0\theta^2}{2} - \frac{\tau n \theta^2}{2} + \tau_0\mu_0\theta + \tau \sum_{i=1}^n X_i\theta\right) \\ &\propto \exp\left[-\frac{1}{2}\left(\theta - \frac{\tau_0\mu_0 + \tau\sum_{i=1}^n X_i}{\tau_0 + n\tau}\right)^2 (\tau_0 + n\tau)\right] \end{aligned}$$

Ainsi $\pi(\theta\,|X_1,\, \ldots,\, X_n)$ est une loi Gaussienne de moyenne $\frac{\tau_0\mu_0 + \tau\sum_{i=1}^n X_i}{\tau_0 + n\tau}$ et de variance $\frac{1}{\sqrt{\tau_0 + n\tau}}$.

On retrouve ainsi la formule que donne IMDb avec les paramètres $\tau = 1$, $v = n$, $C = \mu_0$ et $\tau_0 = m$:

$$\frac{\tau_0\mu_0 + \tau\sum_{i=1}^n X_i}{\tau_0 + n\tau} = \frac{Rv + Cm}{v+m}$$

$$\text{avec :} \qquad R = \frac{1}{n}\sum_i X_i.$$

6.1.2 Agrégation par optimisation

On veut ici voir le classement final comme solution d'un problème d'optimisation combinatoire. Pour cela, nous allons définir des distances entre permutations. On se donne π_i, pour $i \in [\![1,\, p]\!]$ des permutations de $[\![1,\, n]\!]$, et pour une distance d donnée, on souhaite trouver π^* minimisant le problème suivant :

$$\sum_{i=1}^p d(\pi_i,\, \pi^*) \tag{6.1}$$

Dans toute la suite pour $\pi \in \mathfrak{S}_n$, on interprète $\pi(\ell)$ comme la position du candidat ℓ dans l'ordre de préférence.

Définition 6.1.1 (Kendall's τ)
Pour σ et π deux permutations de $[\![1, n]\!]$, on pose :

$$K_{a,b}(\pi, \sigma) = \begin{cases} 1 & si\ \ \pi(a) < \pi(b) \text{ et } \sigma(a) > \sigma(b), \\ 0 & sinon \end{cases}$$

On définit alors la distance $\tau(\pi, \sigma)$ de la manière suivante.

$$\tau(\pi, \sigma) = \sum_{a,b \in [\![1,n]\!]} K_{a,b}(\pi, \sigma)$$

Exemple : Pour $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix}$ et $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 1 & 2 \end{pmatrix}$, on trouve $\tau(\pi, \sigma) = 3$.

Remarque 6.1.1 (majoration de la distance de Kendall)
τ est bien une distance sur $\mathfrak{S}_n$. Par ailleurs, on a toujours :

$$\tau(\pi, \sigma) \leq \frac{n(n-1)}{2}$$

Proposition 6.1.1
Si π^* minimise $\sum_{i=1}^{k} \tau(\pi, \pi_i)$, alors π^* satisfait la condition de Condorcet étendue

Démonstration. On raisonne par l'absurde. On prend X et Y formant une partition non triviale de $[\![1, n]\!]$. On suppose que l'on a $\pi(X) < \pi(Y)$ pour tout i, mais $\pi^*(x) > \pi^*(y)$ pour un couple $x \in X$ et $y \in Y$.

De même que précédemment, on peut trouver x et y vérifiant les mêmes conditions, mais avec $\pi^*(y) = \pi^*(x) - 1$. En inversant les rangs de x et y dans le classement final, on trouve une permutation pour laquelle la grandeur $\sum_i \tau(\pi_i, \pi^*)$ est strictement plus petite. □

Cependant, la minimisation de $\sum_{i=1}^{k} \tau(\pi, \pi_i)$ est NP-difficile dès que $k \geq 4$.

Définition 6.1.2 (Spearman's footrule)
Pour σ et π deux permutations de $[\![1, n]\!]$, on pose :

$$F(\pi, \sigma) = \sum_{i=1}^{n} |\pi(i) - \sigma(i)|$$

Remarque 6.1.2

F est bien une distance sur $\mathfrak{S}_n$. Par ailleurs, on a toujours :

$$F(\pi, \sigma) \leq \left\lfloor \frac{n^2}{2} \right\rfloor$$

Théorème 6.1.1

Pour tout σ et π, deux permutations de $[\![1, n]\!]$, on a les inégalités suivantes :

$$\tau(\pi, \sigma) \leq F(\pi, \sigma) \leq 2\tau(\pi, \sigma)$$

Démonstration.

(1) Montrons la seconde inégalité par récurrence sur la valeur de $\tau(\pi, \sigma)$.

Si $\tau(\pi, \sigma) = 0$, l'inégalité est triviale. τ et F étant des distances, elles sont toutes les deux nulles.

Si $\tau(\pi, \sigma) \geq 1$, on peut construire $\widetilde{\sigma}$ telle que $\tau(\pi, \widetilde{\sigma}) = \tau(\pi, \sigma) - 1$. (Il suffit d'appliquer un pas du tri bulle en triant selon l'ordre de π). On a alors, par inégalité triangulaire : $F(\pi, \sigma) \leq F(\pi, \widetilde{\sigma}) + F(\widetilde{\sigma}, \sigma)$. Et comme $F(\widetilde{\sigma}, \sigma) = 2$ et $F(\pi, \widetilde{\sigma}) \leq 2\tau(\pi, \widetilde{\sigma})$ par hypothèse de récurrence, on trouve bien

$$F(\pi, \sigma) \leq F(\pi, \widetilde{\sigma}) + F(\widetilde{\sigma}, \sigma) \leq 2\tau(\pi, \widetilde{\sigma}) + 2 \leq 2\tau(\pi, \sigma)$$

(2) On peut supposer, sans perte de généralité, que σ est l'identité. Montrons la seconde inégalité par récurrence.

Si $\pi \neq I_d$, on peut considérer i le rang à partir duquel π est égal à l'identité : $\pi(i) < i$ et $\pi(k) = k$ si $k > i$. Il existe alors, par principe de Dirichlet, j maximal vérifiant :

- $\pi(j) > \pi(i)$
- $\pi(j) > j$
- $\pi(i) \geq j$

On constuit alors $\widetilde{\pi}$ à partir de π en échangeant les rangs de i et j.

On a alors : $F(I_d, \pi) = 2\left|\pi(i) - \pi(j)\right| + F(I_d, \widetilde{\pi})$.

Par définition, $\tau(\pi, \widetilde{\pi})$ correpond au nombre d'échanges nécessaires pour passer de $\widetilde{\pi}$ à π. Il suffit d'échanger les places de $\pi(i)$ et $\pi(j)$ ce qui nécessite $2\left|\pi(i) - \pi(j)\right| - 1$ échanges.

Par inégalité triangulaire, on trouve alors :

$$\tau(I_d, \pi) \leq \tau(I_d, \widetilde{\pi}) + \tau(\widetilde{\pi}, \pi) = \tau(I_d, \widetilde{\pi}) + 2\left|\pi(i) - \pi(j)\right| - 1$$

$$\text{Par hypothèse de récurrence} \quad \tau(I_d, \pi) \leq F(I_d, \widetilde{\pi}) + 2\left|\pi(i) - \pi(j)\right| - 1 < F(I_d, \pi)$$

□

On peut montrer que minimiser $\sum_i F(\pi_i, \pi^F)$ se fait en temps polynômial de la manière suivante. On construit le graphe biparti suivant : un sous-ensemble des sommets est

constitué par les candidats et l'autre sous-ensemble par les positions. On considère alors le graphe biparti complet avec cette partition des sommets. Le poids d'une arête entre le candidat j et la position p est alors $\phi_{j,p} = \sum_{i=1}^{k} |p - \pi_i(j)|$.

Pour chaque candidat j, on doit trouver une position $p(j)$, i.e. un matching parfait de telle sorte que $\sum_j \phi_{j,p(j)}$ soit minimisé. Donc le problème d'optimisation est celui de trouver un matching parfait de coût minimum dans le graphe biparti complet où chaque arête a pour coût $\phi_{j,p}$.

Soit π^F, minimisant $\sum_i F(\pi_i, \pi^F)$ et π^* minimisant $\sum_i \tau(\pi^*, \pi_i)$. On a alors

$$\sum_i \tau(\pi^*, \pi_i) \leq \sum_i \tau(\pi^F, \pi_i) \leq \sum_i F(\pi^F, \pi_i) \leq \sum_i F(\pi^*, \pi_i) \leq 2 \sum_i \tau(\pi^*, \pi_i).$$